

François Rebaudo

R pour les scientifiques

Mise en oeuvre de projets
et valorisation des résultats

DUNOD

Illustration de couverture : Login – AdobeStock.com

<p>Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.</p> <p>Le Code de la propriété intellectuelle du 1^{er} juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements</p>	 <p>DANGER LE PHOTOCOPIAGE TUE LE LIVRE</p>	<p>d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée.</p> <p>Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).</p>
--	--	--

© Dunod, 2021

11 rue Paul Bert, 92240 Malakoff

www.dunod.com

ISBN 978-2-10-081547-0

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2^o et 3^o a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

Table des matières

Avant-propos	1
0.1 Pourquoi se former à R	1
0.2 Les objectifs de l'ouvrage	1
0.3 Lectures complémentaires	2
0.4 Remerciements	2
I Les concepts de base	3
1 Premiers pas	5
1.1 L'installation de R	5
1.2 R comme calculatrice	6
1.3 La notion d'objet	13
1.4 Les scripts	15
1.5 En bref	17
2 Choisir un environnement de développement	19
2.1 Éditeurs de texte et environnement de développement	19
2.2 RStudio	20
2.3 Notepad++ avec Npp2R	23
2.4 Geany	26
2.5 Les autres solutions	27
2.6 En bref	27
3 Les types de données	29
3.1 Le type <code>numeric</code>	29
3.2 Le type <code>character</code>	32
3.3 Le type <code>factor</code>	34
3.4 Le type <code>logical</code>	35
3.5 À propos de <code>NA</code>	36
3.6 En bref	37

4	Les conteneurs de données	39
4.1	Le conteneur <code>vector</code>	39
4.2	Le conteneur <code>list</code>	49
4.3	Le conteneur <code>data.frame</code>	62
4.4	Le conteneur <code>matrix</code>	68
4.5	Le conteneur <code>array</code>	74
4.6	En bref	76
5	Les fonctions	77
5.1	Qu'est-ce qu'une fonction	77
5.2	Les fonctions les plus courantes	79
5.3	Les autres fonctions utiles	101
5.4	Écrire une fonction	108
5.5	Les packages	113
5.6	En bref	115
6	Importer et exporter des données	117
6.1	Lire des données depuis un fichier	117
6.2	Exporter ou charger des données pour R	123
6.3	Exporter des données	124
6.4	En bref	125
7	Algorithmique	127
7.1	Les tests logiques avec <code>if</code>	127
7.2	Les tests logiques avec <code>switch</code>	131
7.3	La boucle <code>for</code>	132
7.4	La boucle <code>while</code>	138
7.5	La boucle <code>repeat</code>	139
7.6	<code>next</code> et <code>break</code>	140
7.7	Les boucles de la famille <code>apply</code>	142
7.8	En bref	152
8	Gérer un projet avec R	153
8.1	Gérer des fichiers et des répertoires de travail	153
8.2	Gérer des versions de script	154
8.3	Gérer la documentation	155
8.4	Communiquer avec <code>rmarkdown</code>	157
8.5	En bref	158

II Les graphiques	159
9 Les graphiques simples	161
9.1 La fonction <code>plot</code>	161
9.2 La fonction <code>hist</code>	169
9.3 La fonction <code>barplot</code>	170
9.4 La fonction <code>boxplot</code>	176
9.5 Les autres graphiques	180
9.6 En bref	180
10 La gestion des couleurs	181
10.1 La fonction <code>colors()</code>	182
10.2 La fonction <code>rgb()</code>	184
10.3 Les palettes	185
10.4 En bref	193
11 Les packages graphiques	195
11.1 Les packages de palettes	195
11.2 Le package <code>ggplot2</code>	199
11.3 Les graphiques interactifs et dynamiques avec <code>Plotly</code>	207
11.4 En bref	208
12 Du graphique à la figure dans un article scientifique	209
12.1 <code>Inkscape</code>	210
12.2 <code>The Gimp</code>	211
12.3 Les contraintes techniques de quelques revues	211
12.4 En Bref	211
III Annexes	213
13 Manipuler des dates et des heures	215

Avant-propos

Sommaire

0.1 Pourquoi se former à R	1
0.2 Les objectifs de l'ouvrage	1
0.3 Lectures complémentaires	2
0.4 Remerciements	2

0.1 Pourquoi se former à R

Le logiciel et langage de programmation R s'est imposé comme un outil incontournable d'analyse et de gestion des données scientifiques (et des données d'une manière générale). Il devient indispensable dans ce contexte d'en maîtriser *a minima* les bases. Le succès de R n'est pas un hasard : R est un logiciel que tout le monde peut se procurer librement, et son fonctionnement à base de scripts assure la **transparence** et la **reproductibilité** des résultats scientifiques (sous réserve de respecter quelques règles que nous aborderons dans ce livre). R repose sur une **communauté** très active à l'origine de la création de plusieurs milliers de modules complémentaires (packages) permettant les analyses statistiques les plus pointues. Il est disponible sur les principaux systèmes d'exploitation (Linux, OSX et Windows). Les codes R (ou scripts) sont, sauf cas exceptionnels, indépendants du système d'exploitation utilisé, assurant ainsi leur **portabilité**. R est également un outil qui s'adapte à tous les besoins, depuis de simples statistiques descriptives pour des petits jeux de données jusqu'à la gestion et l'analyse de gros jeux de données (SIG, génomes, BigData ...), que ce soit en local sur un ordinateur, ou à distance sur des serveurs.

0.2 Les objectifs de l'ouvrage

Ce livre est né de la demande des étudiants et professionnels que j'ai eu la chance de rencontrer et de former à R. Les scientifiques, les étudiants et autres personnes souhaitant s'initier à R y trouveront toutes les ressources nécessaires à la mise en œuvre de leurs propres projets scientifiques et à la valorisation de leurs résultats. Il existe de nombreux livres dédiés à R, mais aucun ne couvre les éléments de base de ce langage qui permettent de rendre des résultats scientifiques publiables et reproductibles. Ainsi, tout au long de cet ouvrage nous nous efforcerons de rédiger non seulement un code fonctionnel pour la machine, mais aussi un code lisible et réutilisable pour les humains.

Tous les codes présentés dans ce livre sont disponibles sur dunod.com¹.

0.3 Lectures complémentaires

- “R pour les débutants”, Emmanuel Paradis (https://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf)
- “Introduction à la programmation avec R”, Vincent Goulet (https://cran.r-project.org/doc/contrib/Goulet_introduction_programmation_R.pdf)
- Le blog <http://rzine.fr>, un site collaboratif et interdisciplinaire sur la pratique de R en sciences humaines et sociales

0.4 Remerciements

Je remercie toutes celles et ceux qui ont participé à améliorer ce livre par leurs conseils, leurs suggestions de modifications et leurs corrections (par ordre alphabétique) : Camila BF, Marc G, Susi LH, Emmanuel P, Estefania QH, Baptiste R et Jean-Christophe S.

1. <https://www.dunod.com/>

Première partie

Les concepts de base

Chapitre 1

Premiers pas

Sommaire

1.1 L'installation de R	5
1.2 R comme calculatrice	6
1.2.1 Les opérateurs arithmétiques	6
1.2.2 Les opérateurs de comparaison	8
1.2.3 Les opérateurs logiques	12
1.2.4 Aide sur les opérateurs	13
1.3 La notion d'objet	13
1.4 Les scripts	15
1.4.1 Créer un script et le documenter	16
1.4.2 Exécuter un script	17
1.5 En bref	17

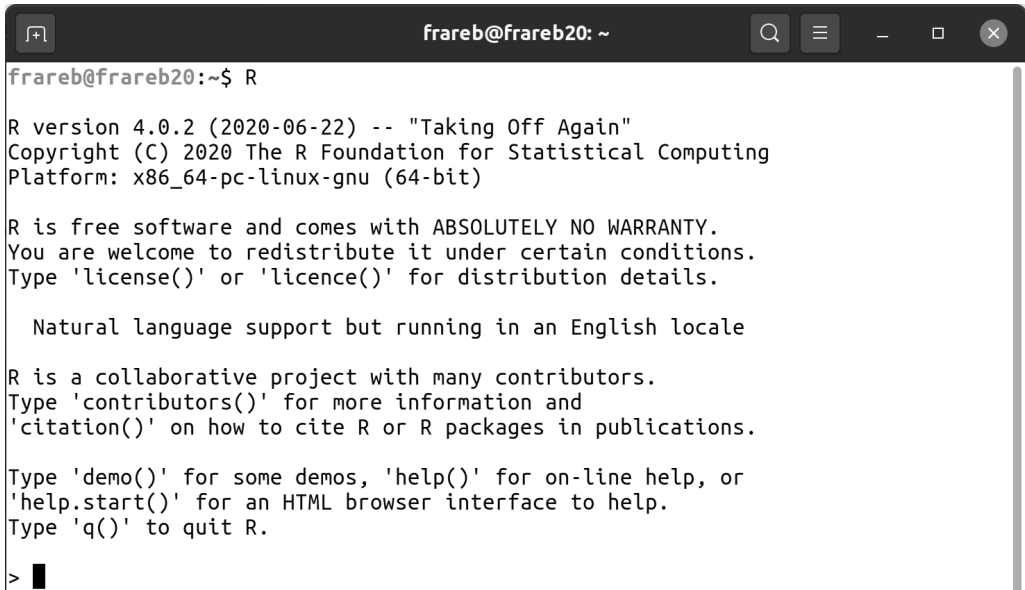
1.1 L'installation de R

Le logiciel R peut être téléchargé depuis de nombreux serveurs du CRAN (Comprehensive R Archive Network) à travers le monde. Ces serveurs s'appellent des "miroirs". Le choix du miroir est manuel.

Le programme permettant l'installation du logiciel R peut être téléchargé depuis le site web de R : <https://www.r-project.org/>. Sur ce site, il faut au préalable choisir un "miroir CRAN" (serveur depuis lequel télécharger R; sauf cas particulier, le plus proche de sa localisation géographique), puis télécharger le fichier **base** correspondant à son système d'exploitation. Les utilisateurs de Linux pourront préférer un `sudo apt-get install r-base` (ou équivalent). Au moment d'écrire ce livre, la version disponible est la 4.0.3 dénommée "Bunny-Wunnies Freak Out".

1.2 R comme calculatrice

Une fois le programme lancé, une fenêtre apparaît dont l'aspect peut varier en fonction de votre système d'exploitation (Figure 1.1). Cette fenêtre est dénommée la **console**. La première information que l'on peut trouver sur la console est la version de R utilisée. Il est recommandé de mettre à jour régulièrement sa version de R afin de bénéficier des dernières fonctionnalités.

A screenshot of a terminal window titled 'frareb@frareb20: ~'. The terminal shows the command 'R' being executed. The output displays the R version (4.0.2), copyright information, platform details (x86_64-pc-linux-gnu), and a welcome message with instructions on how to use R, including commands for license, contributors, help, and quitting. The prompt '>' is visible at the bottom left of the terminal area.

```
frareb@frareb20:~$ R
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> █
```

Figure 1.1 – Capture d'écran de la console R sous Linux Ubuntu.

La console correspond à l'interface sur laquelle va être interprété le code, c'est-à-dire l'endroit depuis lequel le code va être transformé en langage machine, puis exécuté par l'ordinateur. Le résultat de cette exécution sera retransmis dans la console sous une forme lisible par des humains. Cela correspond à l'écran d'affichage d'une calculatrice. C'est de cette manière que R va être utilisé dans la suite de cette section.

Tout au long de ce livre, les exemples de code R apparaîtront sur un fond gris. Ils peuvent être copiés et collés directement dans la console (pour la version électronique de ce livre, et alternativement depuis le site [dunod.com](https://www.dunod.com/)¹), bien qu'il soit préférable de reproduire soi-même les exemples dans la console (ou plus tard dans les scripts). Le résultat de ce qui est envoyé à la console apparaîtra sur fond blanc avec “##” précédant le code afin de distinguer le code et le résultat du code.

1.2.1 Les opérateurs arithmétiques

```
5 + 5
```

1. <https://www.dunod.com/>